有向グラフにおけるノードクラスタリング法とその応用について (函数解析学による一般化エントロピーの新展開)

AUTHOR(S):

保福, 一郎

CITATION:

保福, 一郎. 有向グラフにおけるノードクラスタリング法とその応用について (函数解析学による一般化エントロピーの新展開). 数理解析研究所講究録 2013, 1852: 40-51

ISSUE DATE:

2013-09

URL:

http://hdl.handle.net/2433/195160

数理解析研究所講究録
第1852巻 2013年 40-51

有向グラフにおけるノードクラスタリング法とその応用について

東京都立産業技術高等専門学校・数学教室　保福　一郎
Ichiro Hofuku
Laboratory of Mathematics,
Tokyo Metropolitan College of Industrial Technology

1　はじめに

有向グラフは，与えられた要素(以下，ノードと記す)間の関係を矢線で表現する有効なツールの1つであり，矢線の方向は，ノード間の因果関係あるいは依存関係等により決定される．本研究では，与えられた有向グラフの関係からノード間のクラスタリングを行う手法（以下，ノードクラスタリング法と記す）を提案し，それぞれ特徴をもったノード間のグループ形成を行う．さらにノードクラスタリング法を用いて与えられた有向グラフの構造を解析する新たな手法を提案し，その適用事例を与える．

ここで本研究で用いる用語を紹介し，本研究で適用する有向グラフの条件について述べる．

用語の説明

ノード n_a がノード n_b に対し矢印の向きを与えている場合，ノード n_a はノード n_b にアウトリンクしているといい，逆にノード n_b はノード n_a からインリンクしているという(図1-(i)参照)．またノード n_x が他のノードにアウトリンクしている場合，ノード n_x はアウトリンクを持つといいHubと呼ぶ(図1-(ii)参照)．また逆にノード n_y が他のノードからインリンクしてしている場合，ノード n_y はインリンクを持つといいAuthorityと呼ぶ(図1-(iii)参照)．

条　件 1. (有向グラフの条件) 本研究で扱う有向グラフの任意のノードは，少なくとも *1* つのインリンクもしくはアウトリンクを持つものとする．

条件1は，有向グラフの中で他のノードと関係を持たない，独立に存在するノードは有向グラフのノードから除外することを表している．これらのグラフの矢線構造を基に，ある数理手法を適用すれば有向グラフにおける次の2つの尺度，

(1a) 各ノードの重要度を表す尺度

(1b) 任意の2つのノード間同士の関連性を表す尺度

を導出することができる．項目(1a)についての問題を解決する数理モデルとして，情報検索分野において開発されているWeb順位決定モデルの適用が考えられる．これらのモデルは，各Web

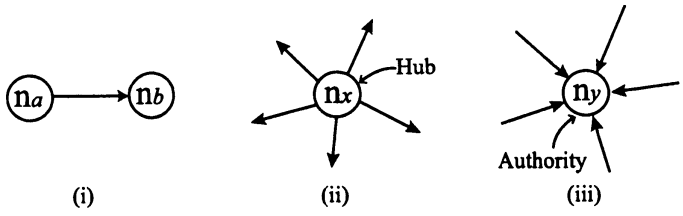

図1: HubとAuthority

ページ間のインリンクとアウトリンクの依存関係を考慮し，重要度の高い順にWebページを決定することが可能である．したがって各Webページをノードに例え，インリンク・アウトリンクの関係をノード間の依存関係を表す矢線で表現すれば，Web内で適用する順位決定モデルは，有向グラフにおけるノード内での重要度を表すことになり(1a)を解決することができる．しかしこれらのモデルは比較的単純なモデルであるため，項目(1b)のノード間の関連性の変化等を解析するまでには至っていない．(1b)の課題を解決するには，グラフのより複雑な構造を表現する新たなモデルが必要となる．そこで著者らはWeb順位決定モデルとして代表的なPagerankとHitsの特性を合わせたアルゴリズム(PH algorithm)を提案することにより項目(1b)の尺度を導出し，ノード間同士のより詳しい関連性を把握することができた[5]．そこで本研究では矢線だけで表現された有向グラフに対し，2つの尺度(1a) (1b)を適用して，次の(A), (B)の課題を解決する手法を提案する．

(A) ノード間同士を，ある特性に合わせたグループに分類する．

(B) 各々のグループの有向グラフに対する影響度を図示化する．

図5, 図6は，それぞれ図4で与えられた有向グラフに対し，項目（A), (B)を行った結果を図示したものである．項目(B)のクラスタリングの生成過程により，グループの位置が下段に位置すれば位置するほと，そのグループで形成するノード間の関係が全体のグラフ構造に与える影響力が強くなり，最下段に位置するグループのノード間の関係が有向グラフの骨格に対応する．以下，項目(A), (B)における解析の概略を与える．

項目(A), (B)の解析の基はPH algorithmの適用による．まず，Authority及びHubのそれぞれに対応したPH algorithmを作成し，有向グラフのノード間のより深い関連性を導出する．次に，この関連性を表す尺度を基に，次に示す2つの項目(1x), (1y)の解析過程を通じて項目(A), (B)を解決するのである．

(1x) 有向グラフにおけるノード間の辺の向きの状況から，入力の中心となるノードの集合，出力の中心となるノードの集合及び入力と出力を繋ぐノードの集合の導出．

(1y) 与えられた有向グラフのノード間のクラスタリング手法の提案

以下第2章にてPH algorithmを紹介し，第3章にて(1x)及び(1y)についての手法を与える．さらに第4章にてノードクラスタリグの適用事例を紹介し第5章で今後の課題について述べる．

2 PH algorithm

本章では本研究にて適用するPH algorithmの紹介を行うが，その前にこのアルゴリズムを適用する際に必要なRanking(I)という1つのランキング手法についての簡単な解説を与える．

2.1 Ranking(I)

本節では，Ranking(I)という1つのランキング手法についての簡単な解説を与える(詳しい解説については文献[4]参照)．ここで，ランキング対象となる要素の集まりを構成集合と呼びCで表す．順位決定における解析の基は，構成集合C内での要素間での一対比較により非負の既約行列を生成し，ベキ乗法を適用するというものである．次に示す条件が，$C=\{c(1),c(2),\cdots,c(n)\}$とした場合の非負既約行列$\mathbf{M}_{(\mathrm{I})}=\{m_{(\mathrm{I})}(i,j)\}_{1\leq i,j\leq n}$の生成法である．

条 件 2. (評価行列Iの生成法)

(2.1a) 行列$\mathbf{M}_{(\mathrm{I})}$は既約でありかつ原始である．

(2.1b) $m_{(\mathrm{I})}(i,j)$は$c(i)$の$c(j)$に対する優位性を示す尺度（優位率と記す）であり，非負の値によって表現される．

(2.1c) 優位率は構成集合 C に属する全ての要素間において，ある定まった共通の規則に基づいて導出される．

条件 2 により生成された行列 $\mathbf{M}_{(\mathrm{I})}$ を 集合 C に対応した評価行列 I と呼ぶ．ここで評価行列 $\mathbf{M}_{(\mathrm{I})}$ に対し，次の反復計算を行う．

$$\left\{\begin{array}{rcl} \mathbf{M}_{(\mathrm{I})}\boldsymbol{u}_{\mathrm{M}_{(\mathrm{I})}[k-1]} & \equiv & \boldsymbol{v}_{\mathrm{M}_{(\mathrm{I})}[k]} \\ \boldsymbol{u}_{\mathrm{M}_{(\mathrm{I})}[k]} & \equiv & \dfrac{\boldsymbol{v}_{\mathrm{M}_{(\mathrm{I})}[k]}}{||\boldsymbol{v}_{\mathrm{M}_{(\mathrm{I})}[k]}||} \qquad k=1,2,\cdots \\ \boldsymbol{p}_{\mathrm{M}_{(\mathrm{I})}[k]} & \equiv & \boldsymbol{u}_{\mathrm{M}_{(\mathrm{I})}[k]} \end{array}\right. \tag{2.1}$$

式 (2.1) における $\boldsymbol{u}_{\mathrm{M}_{(\mathrm{I})}[0]} = {}^{\mathrm{T}}(1,\cdots,1)$ として初期ベクトルと呼び，$\boldsymbol{p}_{\mathrm{M}_{(\mathrm{I})}[k]}$ を行列 $\mathbf{M}_{(\mathrm{I})}$ の k 次ポテンシャルと呼ぶ．また $k=1$ のときの $\boldsymbol{p}_{\mathrm{M}_{(\mathrm{I})}[1]}$ を初期ポテンシャルと呼ぶ．

条件 2 で生成された行列 $\mathbf{M}_{(\mathrm{I})}$ は既約かつ原始であるため，ベキ乗法の適用により行列 $\mathbf{M}_{(\mathrm{I})}$ の絶対値最大な単根である正の固有値 $\lambda_{\mathrm{M}_{(\mathrm{I})}}$ に対応した固有ベクトル $\boldsymbol{r}_{\mathrm{M}_{(\mathrm{I})}}$ を求めることができる (Perron-Frobenius の定理 [2][7])．よって $\lim_{k\to\infty}\boldsymbol{p}_{\mathrm{M}_{(\mathrm{I})}[\mathrm{k}]} \equiv \boldsymbol{p}_{\mathrm{M}_{(\mathrm{I})}[\infty]} = \boldsymbol{r}_{\mathbf{M}_{(\mathrm{I})}}$ と表すことができる．ここで $\boldsymbol{p}_{\mathrm{M}_{(\mathrm{I})}[\infty]}$ を行列 $\mathbf{M}_{(\mathrm{I})}$ の最終ポテンシャルベクトルと呼び，$\boldsymbol{r}_{\mathrm{M}_{(\mathrm{I})}}$ を行列 $\mathbf{M}_{(\mathrm{I})}$ に対応したランキングベクトルと呼ぶ．$\boldsymbol{r}_{\mathrm{M}_{(\mathrm{I})}}$ の生成過程から行列 $\mathbf{M}_{(\mathrm{I})}$ のランキングベクトルについて次の特性を与えることができる．

特 性 1. 行列 $\mathbf{M}_{(\mathrm{I})}$ のランキングベクトル $\boldsymbol{r}_{\mathrm{M}_{(\mathrm{I})}}$ は 1 次ポテンシャルベクトル $\boldsymbol{p}_{\mathrm{M}_{(\mathrm{I})}[1]}$ からの逐次的な推移 (式 (2.1) 参照) により生成された最終ポテンシャルベクトル $\boldsymbol{p}_{\mathrm{M}_{(\mathrm{I})}[\infty]}$ となるため，$\boldsymbol{r}_{\mathrm{M}_{(\mathrm{I})}}$ の各成分に対応する要素 $c(i)$ $(i=1,\cdots,n)$ ではポテンシャルの高い要素に対し，高い優位率を得ている要素の方がランキングが上がる．

ここで $\mathbf{M}_{(\mathrm{I})}$ のランキングベクトルの各成分の大小関係により順位を付けたものを評価行列 $\mathbf{M}_{(\mathrm{I})}$ における集合 C の Ranking(I) と呼ぶ．

2.2 PH algorithm

本節では，Pagerank と Hits を融合した PH algorithm の紹介を行う (詳しい解説については [5] 参照)．まずはじめに，PH algorithm を適用する際に必要な行列 $\mathbf{N} = \{n[i,j]\}$ を以下の法則に従い生成する．

条 件 3. 有向グラフにおけるノード間の関係から行列 $\mathbf{N} = \{n[i,j]\}$ を以下の規則に従い生成する．

$$n[i,j] = \left\{\begin{array}{ll} 1 & \text{ノード } n_i \text{ がノード } n_j \text{ にアウトリンクしている} \\ 0 & otherwise \end{array}\right. \tag{2.2}$$

条件 3 により行列 $\mathbf{N}$ が生成される．この行列を基に PH algorithm を考案する．PH algorithm は次の通りである．

- PH algorithm -

PH 1 : 有向グラフのノード間の関係がら行列 $\mathbf{N}$ を生成し次式により新たな行列 $\mathbf{M} = \{m[i,j]\}$ を生成する．

$$\mathbf{M} = \mathbf{N} + k\,\mathbf{N}^2, \quad 0 \le k \le 1 \tag{2.3}$$

ここで，k はノード n_i からノード n_j へ 2 ステップ目でアウトリンクする矢線の影響度を制御するパラメータである．

PH 2 : 行列 $\mathbf{U}_{\mathrm{A}}, \mathbf{U}_{\mathrm{H}}$ を次の様に定義する.

$$\mathbf{U}_{\mathrm{A}} = {}^{\mathrm{T}}\mathbf{M}\,\mathbf{M}, \qquad \mathbf{U}_{\mathrm{H}} = \mathbf{M}\,{}^{\mathrm{T}}\mathbf{M}$$

ここで, $\mathbf{U}_{\mathrm{A}}$ は Authority に対応した行列を表し, $\mathbf{U}_{\mathrm{H}}$ は Hub に対応した行列を表す.

PH 3 : 行列 $\mathbf{U}_{\mathrm{A}}, \mathbf{U}_{\mathrm{H}}$ の行についてそれぞれ l_1-norm で正規化して行列 $\mathbf{V}_{\mathrm{A}} = \{v_{\mathrm{A}}[i,j]\}$, $\mathbf{V}_{\mathrm{H}} = \{v_{\mathrm{H}}[i,j]\}$ を生成する.

PH 4 : もし行列 $\mathbf{V}_{\mathrm{A}}, \mathbf{V}_{\mathrm{H}}$ の中で行の成分が全て 0 であれば行和が 1 となるような同一の値を与え, 新たな行列 $\mathbf{V}_{\mathrm{A}_1} = \{v_{\mathrm{A}_1}[i,j]\}$, $\mathbf{V}_{\mathrm{H}_1} = \{v_{\mathrm{H}_1}[i,j]\}$ を生成する.

PH 5 : 行列 $\mathbf{V}_{\mathrm{A}_1}, \mathbf{V}_{\mathrm{H}_1}$ の既約性を保証するため, ある調整数 c, $0 < c < 1$ を式 (2.4) の様に作用させ新たに行列 $\mathbf{V}'_{\mathrm{A}_1}, \mathbf{V}'_{\mathrm{H}_1}$ を生成する.

$$\mathbf{V}'_{\mathrm{A}_1} = (1-c)\frac{1}{n}\mathbf{E} + c\mathbf{V}_{\mathrm{A}_1}, \quad \mathbf{V}'_{\mathrm{H}_1} = (1-c)\frac{1}{n}\mathbf{E} + c\mathbf{V}_{\mathrm{H}_1}, \quad 0 < c < 1 \tag{2.4}$$

PH 6 : 次式により行列 $\mathbf{W}_{\mathrm{A}}, \mathbf{W}_{\mathrm{H}}$ をそれぞれ生成する.

$$\mathbf{W}_{\mathrm{A}} = {}^{\mathrm{T}}\mathbf{V}'_{\mathrm{A}_1}, \quad \mathbf{W}_{\mathrm{H}} = {}^{\mathrm{T}}\mathbf{V}'_{\mathrm{H}_1}$$

PH 7 : 行列 $\mathbf{W}_{\mathrm{A}}, \mathbf{W}_{\mathrm{H}}$ のそれぞれの絶対値最大の単根である正の固有値に対する固有ベクトル $r_{\mathrm{W}_{\mathrm{A}}}, r_{\mathrm{W}_{\mathrm{H}}}$ を求める.

2.2.1 ノード間の関連度

本項では, PH algorithm における PH 1 ～ PH 5 を通じ, PH 1 の k に対し, Authority, 及び Hub に関するノード n_i とノード n_j との関連性を示す 1 つの尺度を定義する.

定 義 1. (Authority に関するノード間の関連度) 行列 $\mathbf{V}'_{\mathrm{A}_1}$ の第 i 列を $\boldsymbol{v}'_{\mathrm{A}_1}(i)$, 第 j 列を $\boldsymbol{v}'_{\mathrm{A}_1}(j)$ とする. このとき, $r_{\mathrm{A}}(i,j;k)$ を Authority に関する n_i とノード n_j の関連性を表す尺度 (以下, 関連度と記す) として次の様に定義する.

$$r_{\mathrm{A}}(i,j;k) = \frac{\boldsymbol{v}'_{\mathrm{A}_1}(i) \bullet \boldsymbol{v}'_{\mathrm{A}_1}(j)}{||\boldsymbol{v}'_{\mathrm{A}_1}(i)||_2 \;\; ||\boldsymbol{v}'_{\mathrm{A}_1}(j)||_2} \tag{2.5}$$

(•は内積を表し $||\ ||_2$ は l_2-norm を表す)

定義 1 により生成された $\{r_{\mathrm{A}}(i,j;k)\}$, $(1 \le i,j \le n)$ を $[i,j]$ 成分にもつ行列 $\mathbf{R}_{(\mathrm{A};k)}$ を **Authority に関するノード関連行列**と呼ぶ. 定義 1 により生成された式 (2.5) における $r_{\mathrm{A}}(i,j;k)$ はベクトル $\boldsymbol{v}'_{\mathrm{A}_1}(i)$ とベクトル $\boldsymbol{v}'_{\mathrm{A}_1}(j)$ のなす角を θ とした場合の $\cos\theta$ の値を表す. よってノード n_i とノード n_j との Authority に関する関連度の特性として次の特性を与えることができる.

特 性 2. Authority に関するノード n_i とノード n_j との関連度は, n_i 及び n_j の (自らのノードを含めた) 全ての他のノードに対する関連性の度合いの分布状況の類似性によって決定され, 値の大きい方が関連度が高く, $0 \le r_{\mathrm{A}}(i,j;k) \le 1$ である.

k の値はノード n_i が 2 ステップ目でノード n_j にアウトリンクする矢線の影響度を制御するパラメータである. したがって, k の値が大きくなれば任意の 2 つのノード間の関連度が高くなり, その結果として次の特性が与えられる.

特 性 3. PH 1 の k の値が大きくなればなるほど $r_{\mathrm{A}}[i,j;k]$ は大きくなる.

Hub に関するノード n_i とノード n_j との関連度 $r_{\mathrm{H}}[i,j;k]$ についても Authority の場合と全く同様に次の様に定義することができる.

定　義 2. (**Hub に関するノード間の関連度**) 行列 $\mathbf{V}'_{\mathrm{H}_1}$ の第 i 列を $\boldsymbol{v}'_{\mathrm{H}_1}(i)$ とする．このとき，$r_{\mathrm{H}}[i,j;k]$ を n_i とノード n_j との関連度として次の様に定義する．

$$r_{\mathrm{H}}[i,j;k] = \frac{\boldsymbol{v}'_{\mathrm{H}_1}(i) \bullet \boldsymbol{v}'_{\mathrm{H}_1}(j)}{||\boldsymbol{v}'_{\mathrm{H}_1}(i)||_2 \ ||\boldsymbol{v}'_{\mathrm{H}_1}(j)||_2} \tag{2.6}$$

(•は内積を表し $||\ ||_2$ は l_2-norm を表す)

定義2により生成された $\{r_{\mathrm{H}}[i,j;k]\}$, $(1 \leq i,j \leq n)$ を (i,j) 成分にもつ行列 $\mathbf{R}_{(\mathrm{H};k)}$ を **Hub に関するノード関連行列**と呼ぶ．Authority 同様，Hub に関しても次の 2 つの特性を与えることができる．

特　性 4. Hub に関するノード n_i とノード n_j との関連度は，n_i 及び n_j の（自らのノードを含めた）全ての他のノードに対する関連性の分布状況の類似性によって決定され，値の大きい方が関連度が高く，$0 \leq r_{\mathrm{H}}[i,j;k] \leq 1$ である．

特　性 5. PH 1 の k が大きくなればなるほど $r_{\mathrm{H}}[i,j;k]$ は大きくなる．

PH algorithm の PH 6 における $\mathbf{W}_{\mathrm{A}}$, $\mathbf{W}_{\mathrm{H}}$ の $[i,j]$ 成分は，ノード n_i のノード n_j に対する関連度を表したものである．したがって PH 7 においてべき乗法を用いてそれぞれの行列に対する絶対値最大の固有値に対する固有ベクトルを求める過程は，第 2.1 節における初期ポテンシャルが他のノードに対する関連度の総計ということになり，生成されたランキングベクトルの各成分間には次の特性が与えられる．

特　性 6. PH algorithm を通じて得られたノード間のランキングでは，次の 2 つの特性が与えられる．

(a) 他のノードに対し，高い関連度を示しているノードの方が，そうでないノードよりも順位が上がる．

(b) ポテンシャルが高いノードに対し，より高い関連度を示しているノードの方がそうでないノードよりも順位が上がる．

3　ノードクラスタリング法

本章では，与えられた有向グラフからノード間のクラスタリングを行う手法について解説する．解析の手法としては，まず PH algorithm を基に生成された Authority 及び Hub に関するノードの重要度，及びノード間の関連度から Authority 集合 $\mathrm{D_A}$，及び Hub 集合 $\mathrm{D_H}$ をそれぞれ生成する．次に集合 $\mathrm{D_R} \equiv \mathrm{D_A} \cap \mathrm{D_H}$ とした集合 $\mathrm{D_R}$ (中継集合と記す）を生成し，中継ノードを選定するのである．以下，集合 $\mathrm{D_A}$ 及び $\mathrm{D_H}$ を生成するために必要な確率の導入法について解説する．

3.1　確率の導入

与えられた有向グラフの全てのノードの中から，まず 1 つのノード を選ぶ試行を行い，ノード n_i が選ばれるという事象を A_i とした確率分布 $\mathrm{P}(A_i)$, $(i=1,\cdots,n)$ を考える．次に，1 つめのノードが選ばれたのち，次のノードを選ぶという試行を行い，n_j が選ばれるという事象を B_j とした確率分布 $\mathrm{P}(B_i)$, $(i=1,\cdots,n)$ を考える．この場合，ノード n_j は復元抽出により選ばれたものとする．ここで，選ばれたノード n_i に対し，次に選ばれるノード n_j の条件付き確率

$$\mathrm{P}(B_j|A_i) = \frac{\mathrm{P}(A_i,B_j)}{\mathrm{P}(A_i)} \tag{3.1}$$

を適用し，$\mathrm{D_A}$ 及び $\mathrm{D_H}$ を求めるのである．以下適用法について解説する．

3.2 Authority 集合に関する確率の導入法

本節では，有向グラフの Authority の関係に着目した PH algorithm から導いた行列 $\mathbf{R}_{(\mathrm{A};k)}$ 及びランキングベクトル $\boldsymbol{r}_{\mathrm{W_A}}$ の成分を用い，Authority 集合 $\mathrm{D_A}$ を生成する確率の導入法を解説する．

行列 $\mathbf{R}_{(\mathrm{A};k)}$ の成分 $r_{\mathrm{A}}(i,j;k)$ はノード n_i とノード n_j の関連度を表しているが，ノード n_i に焦点をあてた場合，n_i から見た n_j の関連性をどの様に定義するのかという問題が生じる．実際，ノード n_i が他の各々のノードに対し関連度が高い場合，行列 $\mathbf{W_A}$ のランキングベクトルの生成過程の特徴によりノード n_i の重要度は高くなる特性をもつ（特性 6 参照）．しかし，ノード n_i が他の各々のノードに対し関連度が高ければ高いほど，n_i から相対的に見た他の各々のノードに対する関連性の度合いは低くなると考えられる．そこでノード n_i から見たノード n_j, $(1 \leq j \leq n)$（自らも含む）に対する関連性を表す相対的な指標を**ノード n_i のノード n_j への関係度** (the degree of connection of n_i to n_j) と呼び，次の条件付き確率，

$$\mathrm{P_A}(B_j|A_i) = \frac{1}{\sum_{j=1}^{n} r_{\mathrm{A}}[i,j;k]}\, r_{\mathrm{A}}[i,j;k], \quad (1 \leq i \leq n) \tag{3.2}$$

で定義する．また，式 (3.1) における $\mathrm{P}(A_i)$ については，ノード n_i の重要度が高い方が選ばれる確率が高いものとして，行列 $\mathbf{W_A}$ のランキングベクトル $\boldsymbol{r}_{\mathrm{W_A}} = {}^{\mathrm{T}}(r_{\mathrm{W_A}}(1), \cdots, r_{\mathrm{W_A}}(n))$ に対し，

$$\mathrm{P_A}(A_i) = \frac{1}{\sum_{i=1}^{n} r_{\mathrm{W_A}}(i)}\, r_{\mathrm{W_A}}(i), \quad (1 \leq i \leq n) \tag{3.3}$$

とする．式 (3.2), 式 (3.3) により，$\mathrm{P_A}(A_i, B_j) = \mathrm{P_A}(A_i)\,\mathrm{P_A}(B_j|A_i)$ となり，完全事象系 $\{A_i\}$, $\{B_j\}$ の結合分布を定義することができる．次に $\mathrm{P_A}(A_i, B_j)$ の数理的意味について解説する．

グラフ構造に影響を与える要因の中で，次の 2 つの項目 (3.2a), (3.2b) は大きく影響を与えると考えられる．

(3.2a) 有向グラフの中の重要となるノードの把握

(3.2b) 有向グラフの 2 つのノード間での関連性の把握

$\mathrm{P_A}(A_i)$ はノード n_i の重要度が高ければ高い値を示し，$\mathrm{P_A}(B_j|A_i)$ はノード n_i のノード n_j への関係性の度合いを表す．したがって $\mathrm{P_A}(A_i, B_j)$ の値は，Authority(ノード間のインリンク状況) の観点から生成した，ノード n_i から形成したノード n_j との対グループ ($n_i \circ\!\!\rightarrow n_j$ と記す）の有向グラフの構造に対する影響力の尺度と考えることができる．なぜなら，$\mathrm{P}(A_i)$ の値が高ければ当然そのノード n_i 自身の有向グラフへの影響度は高く (項目 (3.2a) に対応)，またそのノード n_i から形成した関係性の高いノード n_j（項目 (3.2b) に対応）との対グループ $n_i \circ\!\!\rightarrow n_j$ は，当然与えられた有向グラフの構造に対し大きく影響を与えると評価することができるからである．

3.3 Hub 集合に関する確率の導入法

本節では，有向グラフの Hub の関係 (ノード間のアウトリンク状況) に着目した PH algorithm から導いた行列 $\mathbf{R}_{(\mathrm{H};k)}$ 及びランキングベクトル $\boldsymbol{r}_{\mathrm{W_H}}$ を基に，Hub 集合 $\mathrm{D_H}$ を生成する確率の導入法を与える．導入法としては，第 3.2 節で行った解析とまったく同様な手法で完全事情系 $\{A_i\}$, $\{B_j\}$ に対する条件付き確率，$\mathrm{P_H}(B_j|A_i)$, $\mathrm{P_H}(A_i)$ を定義し，$\{A_i\}$, $\{B_j\}$ の結合分布 $\mathrm{P_H}(A_i, B_j) = \mathrm{P_H}(A_i)\,\mathrm{P_H}(B_j|A_i)$ を生成する．以下それぞれの確率を導出する式を与える．

$$\mathrm{P_H}(B_j|A_i) = \frac{1}{\sum_{j=1}^{n} r_{\mathrm{H}}[i,j;k]}\, r_{\mathrm{H}}[i,j;k], \quad (1 \leq j \leq n) \tag{3.4}$$

$$\mathrm{P_H}(A_i) = \frac{1}{\sum_{i=1}^{n} r_{\mathrm{W_H}}(i)}\, r_{\mathrm{W_H}}(i), \quad (1 \leq i \leq n) \tag{3.5}$$

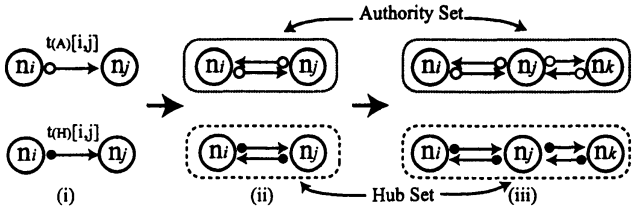

図 2: Authority 集合及び Hub 集合

$\mathrm{P_H}(A_i, B_j)$ の値は，Authority の観点同様，Hub(ノード間のアウトリンク状況) の観点から生成した，ノード n_i から形成した n_j との対グループ ($n_i \bullet\!\!\rightarrow n_j$ と記す) の有向グラフの構造に対する影響力の尺度と考えることができる．

3.4 Authority 集合及び Hub 集合の生成法

ここでは，第 3.2 節にて生成した $\mathrm{P_A}(A_i, B_j)$ 及び第 3.3 節にて生成した $\mathrm{P_H}(A_i, B_j)$ を基に，Authority 集合 $\mathrm{D_A}$ 及び Hub 集合 $\mathrm{D_H}$ の生成法を与える．まずはじめに $\mathrm{P_A}(A_i, B_j)$，$\mathrm{P_H}(A_i, B_j)$ から次に示す確率行列 $\mathbf{T_A}$, $\mathbf{T_H}$ を生成する．

$$\mathbf{T_A} = \{t_\mathrm{A}[i,j]\} = \{\mathrm{P_A}(A_i, B_j)\}, \quad \mathbf{T_H} = \{t_\mathrm{H}[i,j]\} = \{\mathrm{P_H}(A_i, B_j)\} \tag{3.6}$$

式 (3.6) の $\mathbf{T_A}$, $\mathbf{T_H}$ の生成過程の特性により次の式が成立する．

$$\begin{cases} \displaystyle\sum_{i,j=1}^{n} t_\mathrm{A}[i,j] = \sum_{i,j=1}^{n} t_\mathrm{H}[i,j] = 1 \\ \displaystyle\max_{j=1}^{n} t_\mathrm{A}[i,j] = t_\mathrm{A}[i,i], \quad \max_{j=1}^{n} t_\mathrm{H}[i,j] = t_\mathrm{A}[i,i], \quad (i = 1, \cdots, n) \end{cases}$$

以下，$\{t_\mathrm{A}[i,j]\}$, $\{t_\mathrm{H}[i,j]\}$ の各成分を基に，次の規則に従って Authority 集合 $\mathrm{D_A}$ 及び Hub 集合 $\mathrm{D_H}$ を生成する．

- Clustering Algorithm -

CL 1 : (初期段階のクラスタリング) $\{t_\mathrm{A}[i,j]\}$ 及び $\{t_\mathrm{H}[i,j]\}$ の中から値の大きな順に要素を選定し，その要素に対応した 2 つのノード間での関係を図示化する．図示化の方法は，選定した要素が $\{t_\mathrm{A}[i,j]\}$ 及び $\{t_\mathrm{H}[i,j]\}$ のどの要素に対応しているかによって区別をし，次の様に表現する (図 2(i) 参照).

$$\begin{cases} \{t_\mathrm{A}[i,j]\}_{i\neq j}\text{の要素} & \cdots & n_i \circ\!\!\rightarrow n_j \\ \{t_\mathrm{A}[i,i]\}\text{ の要素} & \cdots & n_i \circ \\ \{t_\mathrm{H}[i,j]\}_{i\neq j}\text{の要素} & \cdots & n_i \bullet\!\!\rightarrow n_j \\ \{t_\mathrm{H}[j,j]\}\text{ の要素} & \cdots & n_j \bullet \end{cases}$$

もし，$\{t_\mathrm{A}[i,j]\}$ か $\{t_\mathrm{H}[i,j]\}$ のある値に対するノード間の関係が 2 つ以上存在する場合は，それらの関係の図示化を同時に行う．ここで CL1 の過程において次の 2 つのケースが生じた場合の処理法について示す．

(i) もし 2 つのノード間の関係が互いに向きをもつ状態になったら，図 2(ii) のように Authority 集合 ($\mathrm{D_A}$ と記す) 及び Hub 集合 ($\mathrm{D_H}$ と記す) を形成する．この操作を繰り返していき，図 2(iii) の様に Authority 集合もしくは Hub 集合に属するノードの数を増やしていく．

(ii) もし，2 つ以上の Authority 集合もしくは Hub 集合が存在したら，それぞれの集合を $\mathrm{D_A^{(1)}}, \mathrm{D_A^{(2)}}, \cdots$ または $\mathrm{D_H^{(1)}}, \mathrm{D_H^{(2)}}, \cdots$ の様に，生成された順もしくは更新された順に表す．

ここで，次の条件を満たすノードの名称を与える．

中継ノード ノード n_α が $n_\alpha \in \mathrm{D_A} \cap \mathrm{D_H}$ であるとき，n_α を**中継ノード**と呼びこのノードが属する集合 $\mathrm{D_R} = \mathrm{D_A} \cap \mathrm{D_H}$ を**中継集合**と呼ぶ．

CL 2： 次の (3.4c) または (3.4d) のケースを満たすノード n_y が存在した場合，CL 1 の操作を停止させる (**初期段階のクラスタリング終了**)．

(i) $n_x \in \mathrm{D_A}$ であるノード n_x に対し $n_x \circ\!\!\rightarrow n_y$ であり，かつ $n_y \in \mathrm{D_H} - \mathrm{D_R}$ となるノード n_y が存在した場合

(ii) $n_x \in \mathrm{D_H}$ であるノード n_x に対し，$n_x \bullet\!\!\rightarrow n_y$ であり，かつ $n_y \in \mathrm{D_A} - \mathrm{D_R}$ となるノード n_y が存在した場合

CL 2-(i), (ii) のケースはノード n_x とノード n_y の過剰な関係が生じた状態を意味する．なぜならば，特性的に相反する Authority 集合と Hub 集合に対し，それら 2 つの集合に関係をもたせるノード n_x とノード n_y が存在するということは，与えられたノード間の関係を表現する限度が超えた状況であると判断できるからである．

CL 3： (**第 1 段階のクラスタリング**) CL 2 終了後，再び $\{t_\mathrm{A}[i,j]\}$ 及び $\{t_\mathrm{H}[i,j]\}$ の中から値の大きな順に要素を選定し，その要素に対応した 2 つのノード間での関係を CL 1 と同様の手法で図示する．その際，CL 1〜CL 2 で生成された $\mathrm{D_A}$ 及び $\mathrm{D_H}$ に属するノードに対しては，CL 3 の対象から除外する．

CL 3 により，更に別の Authority 集合あるいは Hub 集合が生成された場合，それぞれ第 1 Authorith 集合 ($\mathrm{D_{A[1]}}$ と記す)，第 1Hub 集合 ($\mathrm{D_{H[1]}}$ と記す) と呼ぶ．また CL 1 同様，ノード n_z が $n_z \in \mathrm{D_{A[1]}} \cap \mathrm{D_{H[1]}}$ を満足すれば n_z を第 1 中継ノードと呼ぶ．このノードが属する集合 $\mathrm{D_{R[1]}} \equiv \mathrm{D_{A[1]}} \cap \mathrm{D_{H[1]}}$ を第 1 中継集合と呼ぶ．

CL 4： Authority 集合，Hub 集合，Relay 集合に属さないノードが，インリンク，もしくはアウトリンクだけをもつノードになるまで CL 1 から CL 3 の作業を繰り返す．

このアルゴリズムを与えられた有向グラフに適用することにより，与えられたノードを様々な集合，$\mathrm{D_A}$, $\mathrm{D_H}$, $\mathrm{D_R}$, $\mathrm{D_{A[1]}}$, $\mathrm{D_{H[1]}}$, $\mathrm{D_{R[1]}} \cdots$, に分類することができる．この様に，与えられたノードから Authority 集合，Hub 集合及び中継集合を生成することを，**ノードクラスタリングを行う**と呼び，その手法を**ノードクラスタリング法**と呼ぶ．次節では，これらの様々な集合を用い，有向グラフから多段階有向グラフを作成する方法について解説を与える．

3.5 多段階有向グラフの作成法

本節では，与えられた有向グラフから図 5 に示されている様な，多段階有向グラフについての作成法とその特性について解説する．多段階有向グラフの作成法としては，初期段階のクラスタリングで生成された $\mathrm{D_A}$, $\mathrm{D_H}$, $\mathrm{D_R}$ に属するノード間の関係を最も下段 (Bottom stage) に位置させる．これらの集合に属するノードはノード自身の持つ影響度も大きく，ノード間同士の関係性も高い値を示していることから，有向グラフ全体の構造に与える影響力が大きい．よって Bottom stage に位置するノード間同士の関係は，与えられた有向グラフの骨格と見なすことができる．次に $\mathrm{D_{A[1]}}$, $\mathrm{D_{H[1]}}$, $\mathrm{D_{R[1]}}$ に属するノード間の関係については，Bottom stage から 1 つ上の段 (First stage) に位置させる．この様な操作を続けていくことにより有向グラフから多段階有向グラフを作成することができる．多段階有向グラフ生成の特性上，次の特性を与えることができる．

特 性 7. (多段階有向グラフの特性)

(a) 多段に分かれたノード間の関係が下段に位置すれば位置するほど，その段に対応するノード間同士の関係の，全体の有向グラフ構造への影響度は大きくなる．

(b) Bottom stage に位置するノード間の関係は，全体の有向グラフ構造の骨格と見なすことができる．

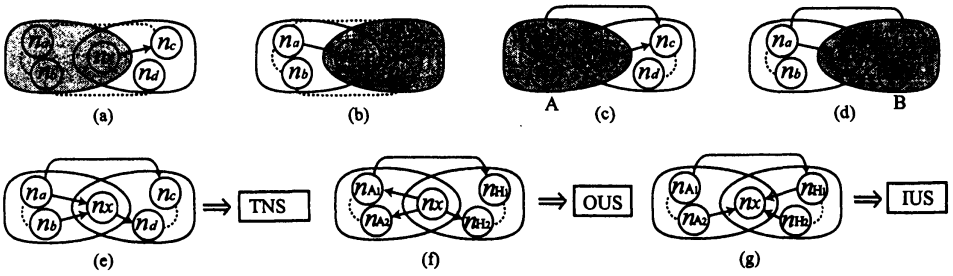

図 3: 改良有向グラフにおける中継ノードの扱い方

3.6 多段階有向グラフから改良有向グラフへの変換

本節では，第 3.5 節で形成した多段階有向グラフから二次元の有向グラフ (改良有向グラフと呼ぶ) への変換の仕方について述べる．変換の基としては，多段階有向グラフにて存在するそれぞれの集合 D_A, D_H, D_R に区別をつけて表現すればよいのだが，ここで中継集合をどの様に改良有向グラフにて表現するかという問題がある．そこで本研究では，与えられた多段階有向グラフから改良有向グラフを作成する場合，中継集合については次の規則により扱う．

規 則 1. 改良有向グラフの中継集合の扱い方を次に定める．ここで中継集合 D_R に対し，ノード n_x を，$n_x \in D_R$ とする．

- 中継集合の扱い方 -

TR 1 : 中継集合 D_R が $D_R = D_A = D_H$ であれば D_R としてそのまま図示化する．

TR 2 : ノード n_x が唯一のアウトリンクしか持たない場合，n_x に関する矢線の始点側に属する集合に n_x を吸収させる (図 3-(a) 参照)．

TR 3 : ノード n_x が唯一のインリンクしか持たない場合，n_x に関する矢線の終点側に属する集合に n_x を吸収させる (図 3-(b) 参照)．

TR 4 : ノード n_x がインリンクとアウトリンクの双方を持つ場合は，次の 2 つのケースを考える．

(i) n_x にアウトリンクする矢線の始点側の集合 A に属する全てのノードが，n_x もしくは相反する集合 (中継集合を省く) のノードにアウトリンクする場合，n_x を集合 A に吸収させる (図 3-(c) 参照)．

(ii) n_x からインリンクする矢線の終点側の集合 B に属する全てのノードが，n_x もしくは相反する (中継集合を省く) 集合のノードからインリンクする場合，n_x を集合 B に吸収させる (図 3-(d) 参照)．

(iii) (i), (ii) の双方を満足する場合，D_A, n_x, D_H を 1 つにまとめ，推移結合集合 (Transitive union set : TNS) と呼び，図 3-(e) で表す．

TR 5 : ノード n_x が D_H, D_A に対しアウトリンクをもち，それぞれの集合に対して TR 4-(i) を満足するとき，D_A, n_x, D_H を 1 つにまとめ，出力結合集合 (Output union set : OUS) と呼び，図 3-(f) で表す．

TR 6 : ノード n_x が D_H, D_A からインリンクをもち，それぞれの集合に対して TR 4-(ii) を満足するとき，D_A, n_x, D_H を 1 つにまとめ，入力結合集合 (Input union set : IUS) と呼び，図 3-(g) で表す．

TR 7 : 上記 (TR 1)～(TR 6) のケースに当てはまらない場合は，n_x を D_A, D_H から独立させて図示化する．

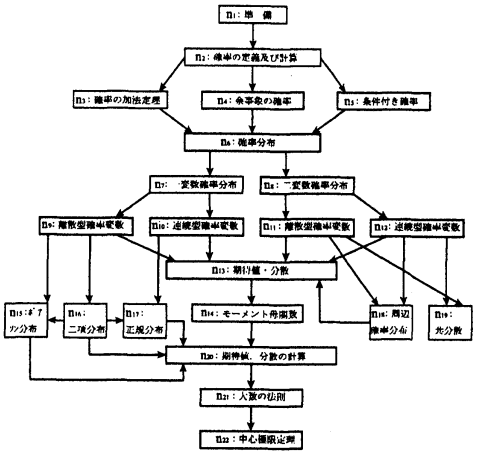

図 4: テキストシラバスから導出した教授項目間の依存関係

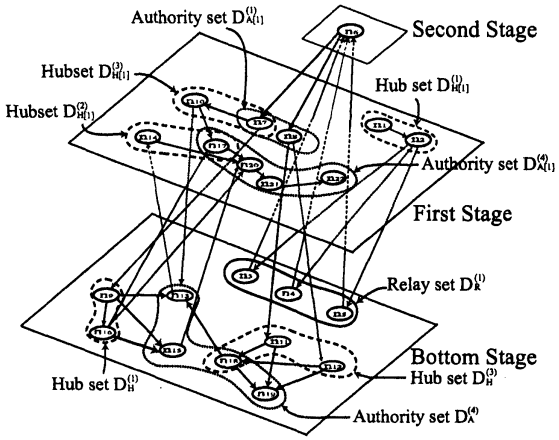

図 5: ノードクラスタリング法を用いた多段階有向グラフ

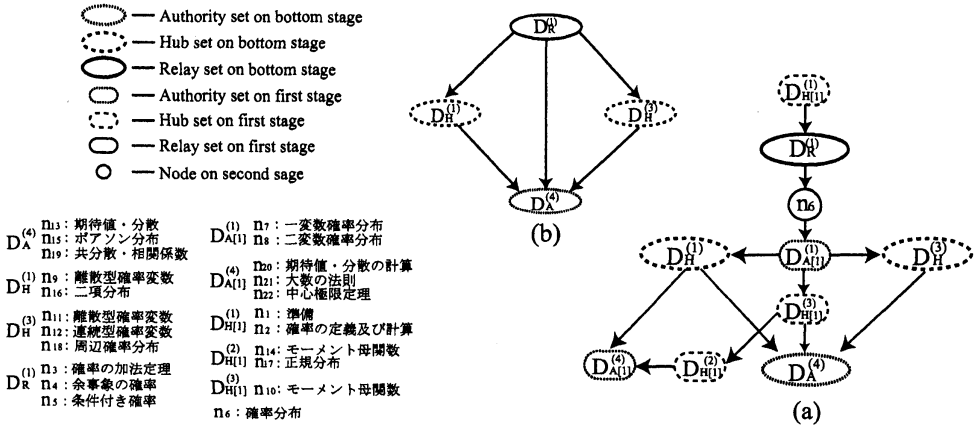

図 6: グラフィックシラバスと有向グラフの骨組み

4 ノードクラスタリング法の適用例

ここ 10 年，大学等の授業では，教員は授業を行う前に授業計画としてシラバスを作成している．シラバスは学生に対し講義内容全体を紹介する上で重要ではあるが，その殆どが文章でかかれている (テキストシラバス) ため，教授項目の関連性やその教科全体の教授項目のイメージを把握することは難しい．この様な問題点を解決する方法として，「授業のおける教授項目間の関連性を図示化する流れ図」を提示するグラフィック・シラバス [8] と呼ばれるものが存在する．グラフィックシラバスの中には，単に教える教授項目の流れを示すだけでなく，教える教科のイメージを独創的なグラフィックを用いて表現しているケースもある．そこで本研究では，著者らが与えた実際のテキストシラバスの教授項目の関係から有向グラフを生成し，ノードクラスタリング法を用いてグラフィックシラバスの生成を試みる．

表 1: 各々のステージで生成された集合

Stage	Hub set	Authority set	Relay set
Bottom stage	$D_H^{(1)}$, $D_H^{(3)}$	$D_A^{(4)}$	$D_R^{(1)}$, $D_R^{(2)}$
First stage	$D_{H[1]}^{(1)}$, $D_{H[1]}^{(2)}$, $D_{H[1]}^{(3)}$	$D_{A[1]}^{(1)}$, $D_{A[1]}^{(4)}$	$D_{R[1]}^{(1)}$, $D_{R[1]}^{(2)}$
Second stage	-	-	-

4.1 適用事例

図 4 は，著者が実際に行っている「確率の授業」におけるテキストシラバスの教授項目の関連性を独自の判断で有向グラフで表したものである．このグラフを基に，ノードクラスタリング法を用いて与えられたノード間でのクラスタリングを行う (式 (2.3) の $k = 0.5$ とした)．クラスタリングの過程であるが，初期段階のクラスタリングは，88 ステップで終了し，第 1 段階のクラスタリングは 38 ステップで終了した．その時点で，残りの対象となるノードが n_6 の 1 つだけとなったので，この段階で有向グラフのクラスタリングが終了した．表 1 は，それぞれの段階で最終的に生成されたそれぞれの集合を記したものである．また図 5 は，ノードクラスタリングの結果を多段階有向グラフの作成法 (3.5) に基づき作成したものである．図 5 を参照すれば解るように生成された多段階有向グラフは 3 つの段階で形成されていることが解る．

図 5 の中継ノード n_{18}, n_{17}, n_7 に対し，規則 1 を適用すると，各々のノードは，

$$n_{18} \in D_H^{(3)}, \quad n_{17} \in D_{H[1]}^{(2)}, \quad n_7 \in D_{A[1]}^{(1)}$$

の様に，Authority 集合，及び Hub 集合に吸収される．その結果を基に改良有向グラフを生成すると図 6-(a) となった．この図 6-(a) がノードクラスタリング法を適用した場合の「確率の授業」におけるグラフィック・シラバスとなる．

4.1.1 ノードクラスタリング法により生成したグラフィック・シラバスの利点

ノードクラスタリング法により生成されたグラフィックシラバスを用い，次の項目を導出することができる．

(4a) 教科全体の講義イメージの骨格

(4b) 重要度の低い教授項目の導出

以下，(4a)，(4b) に対応する関係を説明する．図 6-(b) は教科内容の各単元同士の大域的な関係を講義イメージとして表したものである．このグラフはダイヤモンドの形をしており，教科全体の講義イメージを表す骨格（項目 (4a) に対応）としては極めて整った形となる．表 1 を参照すると，Bottom stage に存在する $D_A^{(4)}$ の教授項目群が，学生が習得すべき最重要項目を表し，$D_H^{(1)}$, $D_H^{(3)}$ の教授項目群が，教える側が学生に対し，導入項目としてしっかりと理解を求める最重要項目を表すことになる．また $D_R^{(1)}$ は，本講義の繋ぎとして極めて重要な教授項目群を表す．繋ぎに含まれる教授項目を理解できなければ，$D_H^{(1)}$, $D_H^{(3)}$ の教授項目群が理解できたとしても，$D_A^{(4)}$ の理解へは当然到達できないことになる．また，図 4 のシラバスによる有向グラフでは，授業目的の最終目標は，n_{22} : 中心極限定理 となるが，ノードクラスタリング法によるグラフィック・シラバスでは，学生が習得すべき最重要項目は $D_A^{(4)}$ の期待値・分散，ポアソン分布，周辺確率分布，共分散・相関係数となっており，確率統計を利用する道具としての内容が多く含まれていることが解る．

5 今後の課題

本研究で提案したグラフ構造の解析は，ノード間に何らかの関係をもった有向グラフであれば全てに対応できるものである．適用事例として著者が実際に行っている「確率の授業」におけるテキストシラバスの教授項目の関連性を有向グラフで表したものからグラフィック・シラバスを作成し，著者の解釈を与えた．

別の観点の適用を考えれば，例えばある学科で開講している各教科のテキストシラバスから，教科間全体での依存関係を表した有向グラフを考える．この有向グラフに対してノードクラスタリング法を適用し，多段階有向グラフを生成すれば，Hub集合，Authority集合，中継集合が導出され，これらの結果から，講義科目が必修科目であるべきか，あるいは，必修選択か，選択科目であるべきかという科目種別決定の1つの根拠となる．またBottom stageに対応する科目群の構造からは，その学科のカリキュラムポリシーを与える1つのフレームができあがることになる．また別の例では，インターネットのページ間のリンク構造が考えられる．現在のインターネットでのWeb情報検索エンジンはページ間の有向グラフの関係を表した巨大なネットワーク構造が土台となっている．これらのページ間の関連性には様々な構造が存在し，その関連性にも当然強弱の量が存在する．これらのネットワーク構造に対し，本手法を適用して新たなネットワーク構造を構築するということは，Web情報検索システム自体の更なる発展に繋がる可能性があると考える．

参考文献

[1] Amy, N.L. and Carl, D.M., *A Survey of Eigenvector Methods for Web Information Retrieval*, SIAM Review, Vol.47, No.1, 2005, pp.135-161.

[2] Berman, A. and Plemmons, R.J., *Nonnegtive Matrices in the Mathematical Science*, Academic Press, New York, 1979.

[3] Hofuku, I. and Oshima, K., *Rankings Schemes for Various Aspects Based on Perron Frobenius Theorem*, INFORMATION, Vol.9, No.1, 2006, pp.37-52.

[4] Hofuku, I. and Oshima, K., *A mathematical structure of processes for generating rankings through the use of nonnegative irreducible matrices*, Applied Mathematics and Information Science, Vol.4, No.1, pp.125-139.

[5] Hofuku.I, Yokoi.T and Oshima.K, *Measures to Represent the Properties of Nodes in a Directed Graph*, INFORMATION, Vol.13, No.3(A), pp.537-549, 2010.

[6] 保福一郎，大島邦夫，相交えない二群間の共通の試技・競技による混合ランキングの生成法について，日本応用数理学会論文誌，Vol.13, No.1, 2003, pp.97-114.

[7] Lancaster, P. and Tismenetsky, M., *The Theory of Matrices*, Academic Press, New York, 1985.

[8] Nilson, Linda Burzotta, *The Graphic Syllabus and the Outcomes Map : Communicating Your Course*, Jossey-Bass Inc Pub.

[9] Ortega, J.M., *Numerical Analysis : A Second Course*, SIAM, Philadelphia, 1990.

[10] 大島邦夫, 保福一郎, ランキングベクトルとウェイトを適用した試験結果におけるランキング法について, 日本応用数理学会論文誌, Vol. 6, No.1, 1996, pp.133-146.